

## BIROn - Birkbeck Institutional Research Online

Studer, R. A. and Christin, P.-A. and Williams, Mark A. and Orengo, C.A. (2014) Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America* 111 (6), pp. 2223-2228. ISSN 0027-8424.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/9096/>

*Usage Guidelines:*

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>  
contact [lib-eprints@bbk.ac.uk](mailto:lib-eprints@bbk.ac.uk).

or alternatively

# Stability-activity tradeoffs constrain the adaptive evolution of RubisCO

Romain A. Studer<sup>a,1</sup>, Pascal-Antoine Christin<sup>b</sup>, Mark A. Williams<sup>c</sup>, and Christine A. Orengo<sup>a</sup>

<sup>a</sup>Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London WC1E 6BT, United Kingdom; <sup>b</sup>Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom; and <sup>c</sup>Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, London WC1E 7HX, United Kingdom

Edited by George H. Lorimer, University of Maryland, College Park, MD, and approved January 2, 2014 (received for review June 6, 2013)

A well-known case of evolutionary adaptation is that of ribulose-1,5-bisphosphate carboxylase (RubisCO), the enzyme responsible for fixation of CO<sub>2</sub> during photosynthesis. Although the majority of plants use the ancestral C<sub>3</sub> photosynthetic pathway, many flowering plants have evolved a derived pathway named C<sub>4</sub> photosynthesis. The latter concentrates CO<sub>2</sub>, and C<sub>4</sub> RubisCOs consequently have lower specificity for, and faster turnover of, CO<sub>2</sub>. The C<sub>4</sub> forms result from convergent evolution in multiple clades, with substitutions at a small number of sites under positive selection. To understand the physical constraints on these evolutionary changes, we reconstructed *in silico* ancestral sequences and 3D structures of RubisCO from a large group of related C<sub>3</sub> and C<sub>4</sub> species. We were able to precisely track their past evolutionary trajectories, identify mutations on each branch of the phylogeny, and evaluate their stability effect. We show that RubisCO evolution has been constrained by stability-activity tradeoffs similar in character to those previously identified in laboratory-based experiments. The C<sub>4</sub> properties require a subset of several ancestral destabilizing mutations, which from their location in the structure are inferred to mainly be involved in enhancing conformational flexibility of the open-closed transition in the catalytic cycle. These mutations are near, but not in, the active site or at intersubunit interfaces. The C<sub>3</sub> to C<sub>4</sub> transition is preceded by a sustained period in which stability of the enzyme is increased, creating the capacity to accept the functionally necessary destabilizing mutations, and is immediately followed by compensatory mutations that restore global stability.

The adaptive diversification of organisms often requires the evolution of novel enzymatic properties. The evolutionary shift from one enzymatic function to another involves crossing an energetic barrier in a fitness landscape (1). The number of mutations that confer advantageous function during such a shift is consequently limited. Some residues are critical for maintaining the stability of the protein fold, others are important for the catalytic activity itself. Due to the multiple roles of amino acids in proteins, the adaptation of one physical parameter of an enzyme is likely to affect other properties (2). As proteins usually form thermodynamically stable structures, their evolutionary trajectories are constrained to a narrow range of stability (3). Stability and activity are likely to be negatively correlated. Most possible amino acid changes in native proteins are destabilizing and, consequently, mutations that lead to a more favorable enzyme activity are likely to decrease the stability of the protein (2, 4). Compensatory mutations are then needed to restore global stability. These processes are referred to as stability-activity tradeoffs (5–7). Furthermore, proteins with higher stability confer greater evolvability, because there is more scope to accept destabilizing yet functionally beneficial changes (8). Whereas such stability activity tradeoffs are well attested in laboratory experiments, it remains unclear as to how strong a signal these particular physical constraints would leave in a naturally, and slowly, evolving population where there are many potentially competing evolutionary pressures and considerable neutral drift (9).

The probability that a new mutation becomes fixed in a species is determined by the relative strengths of genetic drift and natural

selection. Although the rate of fixation is assumed to be constant under neutral evolution, it is decelerated by negative selection, which tends to remove deleterious mutations, or accelerated by positive selection, under which favorable mutations, e.g., those enabling adaptation of the protein following environmental changes, tend to be retained. A well-known case of adaptation under positive selection is ribulose-1,5-bisphosphate carboxylase (RubisCO; Enzyme Commission no. 4.1.1.39), the enzyme responsible for fixation of CO<sub>2</sub> to ribulose-1,5-bisphosphate in the Calvin–Benson cycle. It is the most abundant protein on earth and represents up to 30% of all soluble proteins in plants. However, this abundant enzyme also has a very low turnover of <10/s. RubisCO can catalyze reactions with both CO<sub>2</sub> and O<sub>2</sub>, and the catalytic rate for CO<sub>2</sub> fixation is negatively correlated with CO<sub>2</sub>/O<sub>2</sub> specificity (10). The fixation of O<sub>2</sub> initiates the photorespiratory cycle, which uses ATP to regenerate CO<sub>2</sub>, resulting in both energy loss and a net loss of fixed CO<sub>2</sub>. Because these losses are disadvantageous, there is selection for increased affinity for CO<sub>2</sub> compared with O<sub>2</sub> and thus for low catalytic rates (10). The dual affinity seems inevitable, as both CO<sub>2</sub> and O<sub>2</sub> can attack the carbanion form of ribulose-1,5-bisphosphate produced during the reaction (11).

Several lineages of flowering plants (angiosperms) have evolved mechanisms that diminish photorespiration by concentrating CO<sub>2</sub> before its fixation by RubisCO. These mechanisms operate in various pathways such as crassulacean acid metabolism (CAM) and C<sub>4</sub> photosynthesis. Although CAM is mainly an adaptation to water stress, C<sub>4</sub> photosynthesis is advantageous in all conditions that promote photorespiration, such as warm, open, dry, saline, or some aquatic environments. In C<sub>4</sub> plants, atmospheric CO<sub>2</sub> is initially incorporated into small organic compounds by a series of

## Significance

How enzymes acquire new functions is a key question in evolutionary biology. Here, we studied the evolution of some forms of ribulose-1,5-bisphosphate carboxylase, the enzyme responsible for CO<sub>2</sub> fixation in photosynthesis, which has evolved enhanced activity in multiple groups of plants. We showed that the evolution of this enzyme was constrained by tradeoffs between activity and stability, two key properties of enzymes. The acquisition of enhanced activity was mediated by mutations destabilizing the structure. However, these were preceded and followed by periods in which stabilizing mutations were predominant, so that global stability was always maintained. This work shows that the natural evolution of enzymes is subject to strong biophysical constraints, and evolution follows perilous paths toward adaptation.

Author contributions: R.A.S., P.-A.C., M.A.W., and C.A.O. designed research; R.A.S. performed research; R.A.S., P.-A.C., M.A.W., and C.A.O. analyzed data; and R.A.S., P.-A.C., M.A.W., and C.A.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence should be addressed. E-mail: r.studer@ucl.ac.uk.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310811111/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310811111/-DCSupplemental).

enzymes beginning with carbonic anhydrase and phosphoenolpyruvate carboxylase, a system without affinity for  $O_2$ . These compounds are transported to the specialized compartments (most often distinct cells) where RubisCO is located. The various pathways lead to the formation of malate or oxaloacetate, which are decarboxylated to yield  $CO_2$  and pyruvate or phosphoenolpyruvate (12), producing an up to 10-fold increase of  $CO_2$  concentration in the proximity of RubisCO. Despite its relative complexity, the  $C_4$  trait has evolved more than 62 times in different groups of flowering plants (13), including up to 24 times in grasses alone (14).

The turnover rate of RubisCO is positively correlated with its  $CO_2$  affinity [ $K_m(CO_2)$ ] and negatively correlated with the  $CO_2/O_2$  specificity ratio of the enzyme (10, 15, 16). The high concentration of  $CO_2$  at the site of RubisCO in  $C_4$  plants allows a lower specificity ratio of  $CO_2/O_2$  and therefore an increase in turnover rate and thus efficiency (17, 18). Experimental studies of RubisCOs from very closely related  $C_3$  and  $C_4$  species within the *Flaveria*, *Atriplex*, and *Neurachne* genera showed that very few changes may be necessary to modify enzymatic properties in response to the modification of the metabolic context (19, 20). Indeed, in the *Flaveria* context, a single mutation (M309I) has been identified as key in modifying specificity and increasing turnover (21); it remains unclear as to how this observation applies to a wider range of plants and what the contributions are of other observed mutations to adaptation. Comparative sequence analysis of a broader range of plant species does suggest that, in general, adaptation of RubisCO to  $C_4$  metabolism involves a larger number of amino acid changes found to be under positive selection (19, 20). Here, we investigate the role of mutations in the adaptation of a large group of plants, focusing in particular on the constraints imposed by stability requirements, which have been previously shown to be important in the directed evolution of enzymes.

In this study, we focused on the RubisCO of the monocot lineage, which is one of the major groups of flowering plants and contains both  $C_3$  and  $C_4$  species. Its diversification probably started 120 Mya, and the emergence of distinct  $C_4$  species has occurred over the last 40 My. We took advantage of the convergent nature of the evolution of  $C_4$  photosynthetic pathways and the resulting common changes in the selective pressures on RubisCO to investigate the structural factors influencing the evolvability of novel enzymatic properties. Our combined phylogenetic framework and structural analyses allowed an *in silico* reconstruction of the ancestral sequences and 3D structures of the large subunit within the RubisCO complex. Our investigations have enabled the inference of the mutational paths linked to the adaptation to  $C_4$  photosynthesis in the monocots.

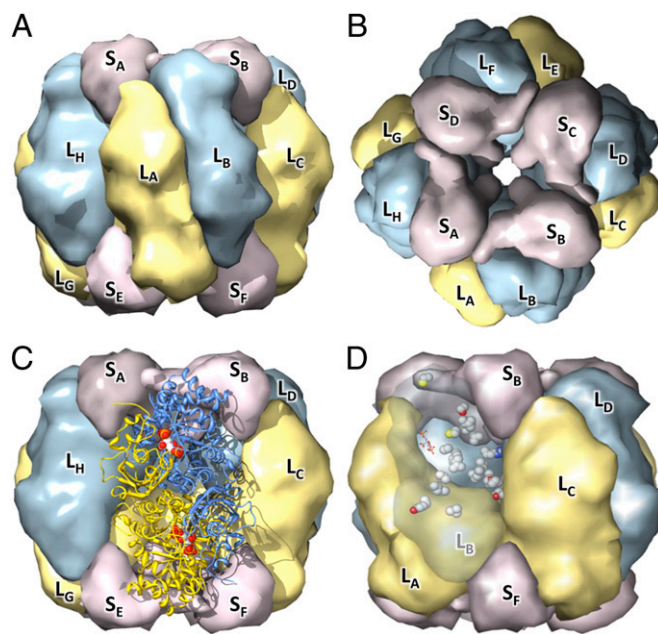
This work shows that the evolutionary adaptation of the RubisCO enzyme is mediated by stability-activity tradeoffs with many stabilizing mutations apparently being fixed simply to allow functionally necessary destabilizing mutations to be tolerated. The enzyme has used multiple paths to adapt to new environmental conditions with no single mutation present in more than two-thirds of  $C_4$  species. The paths are structurally diverse, including the mutation of residues close to and remote from the active site. The location of many of the positively selected mutations implies that allosteric modulation of structure at the active site and (possibly cooperative) dynamics of domain and subunit movements are keys to adaptation.

## Results

**Overview.** The RubisCO of plants, as exemplified by the enzyme from the rice *Oryza sativa*, is a hexadecamer composed of eight large subunits (encoded by the ribulose-bisphosphate carboxylase gene *rbcL*) and eight small subunits (encoded by *rbcS*;  $L_8S_8$ ) (Fig. 1). The following analysis is necessarily limited to the catalytic *rbcL*, because insufficient sequences of monocot *rbcS* genes are available to reliably reconstruct ancestral sequences (see *SI Text* for additional remarks). We divided our analysis of *rbcL* into two parts. First, the stability landscape was investigated by computationally scanning all possible mutations of the *O. sativa*

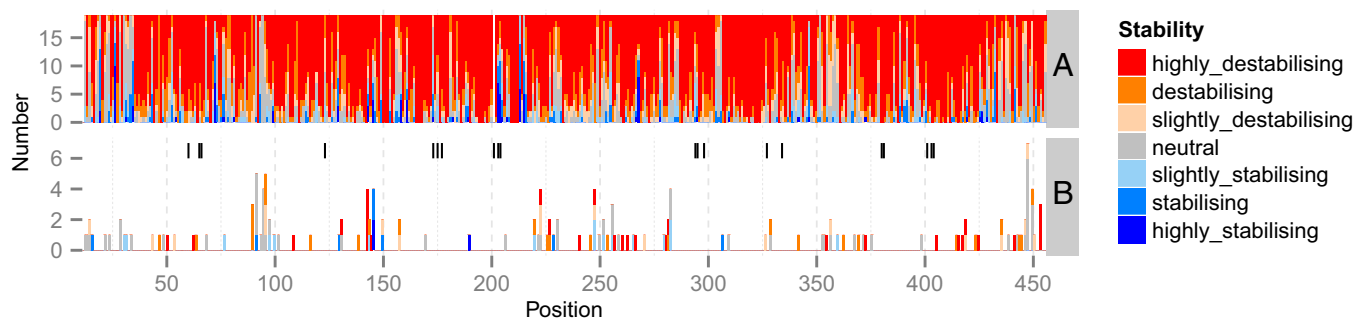
RubisCO, which is a  $C_3$  form (no structure of a RubisCO from a  $C_4$  plant has been determined). Second, ancestral mutations that occurred during the adaptation of RubisCO in monocots were identified, selective pressures were estimated, and the effect of the positively selected mutations on stability and their locations in the 3D structure were examined.

**Stability Landscape of All Possible Mutations.** The stability effect of all possible mutations of each residue of the quaternary complex was estimated using FoldX. The WT amino acid at each position of the *O. sativa rbcL* was mutated (in all eight chains) to each of the 19 other possibilities. This energetic landscape highlights positions that are mutation tolerant (Fig. 2A). For convenience, if we categorize the calculated effects of mutations in proportion to the known accuracy of FoldX predictions (*Methods*), then most possible mutations (5,007 of 8,436 = 59.4%) are found to be highly destabilizing ( $\Delta\Delta G_{fold}$  per chain > +1.84 kcal/mol) and 3,335 mutations (39.5%) have a moderate effect ( $-1.84 < \Delta\Delta G_{fold} < +1.84$  kcal/mol). Ninety-four mutations (1.1%) can strongly stabilize the structure, but only in a smaller number of positions (35/444), most of which are in the active site. It has previously been observed that residues close to an active site are often intrinsically destabilizing, because their great functional utility is traded against stability (22, 23). Finally, less than one quarter of the positions (103/444) were found to be actually mutated in our monocot sequence dataset, with only two to four alternative residues observed at each position (Fig. 2B).



**Fig. 1.** The RubisCO hexadecamer structure. Pairs of large subunits (blue and yellow) form dimers with an extensive interface; four of these dimers form an octameric ring. The interdimer interfaces are comparatively small, and the overall structure is stabilized by the binding of eight small subunits (lavender) that bridge dimers. (A and B) Surface views from side and top, respectively. (C) The two chains forming the  $L_A L_B$  dimer are shown in ribbon form. Each dimer forms two active sites, the upper site here being between the N-terminal domain of  $L_A$  and the C-terminal domain of  $L_B$ . Each site undergoes an open to closed structural transition on substrate binding. The reaction intermediate analog 2-carboxyarabinitol-1,5-bisphosphate is shown bound at each site in this structure (PDB code: 1WDD). The larger C-terminal domain contributes most residues to each active site, but the N-terminal domain is critical for positioning the  $CO_2$  or  $O_2$  molecule. (D) Atoms of residues under positive selection in the large subunit ( $L_8$ ) are shown as spheres. These residues are frequently close to subunit interfaces.





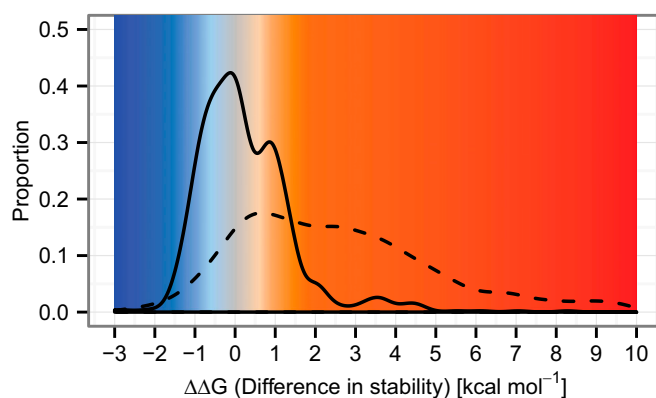
**Fig. 2.** Effect of mutations on protein stability. (A) Stability landscape of the large subunit (*rbcl*). All 19 possible mutations at each position observed in the *O. sativa* structure (positions 12–456) are colored on a vertical bar in terms of their stability relative to the native residue. Residues that are part of the active site are indicated by a black bar. The thresholds for  $\Delta\Delta G_{\text{fold}}$  in kcal/mol are highly stabilizing ( $< -1.84$ ), stabilizing ( $-1.84$  to  $-0.92$ ), slightly stabilizing ( $-0.92$  to  $-0.46$ ), neutral ( $-0.46$  to  $+0.46$ ), slightly destabilizing ( $+0.46$  to  $+0.92$ ), destabilizing ( $+0.92$  to  $+1.84$ ), and highly destabilizing ( $> +1.84$ ). Positions where the vertical bar is substantially gray or blue are predicted to be tolerant of mutation and where largely red are intolerant. Highly destabilizing mutations are very unlikely to occur in nature. (B) Stability effect of observed mutations at each position, relative to the *O. sativa* *rbcl* sequence. Within the monocot species, 105 positions of the 444 aligned residues of the peptide chain have alternate amino acids. The overwhelming majority of observed mutations (79.5%) have modest stability changes in the range of  $-1.84$  to  $+1.84$  kcal/mol.

**Analysis of Mutations Occurring During Evolution and Their Effect on Stability.** The monocot dataset exhibits a  $>95\%$  pairwise sequence identity at the protein sequence level and no alignment gaps. This high level of conservation, together with the previously determined, highly resolved, phylogenetic tree (24), allowed the reconstruction, with high confidence, of the ancestral sequences (each comprising 444 mutable amino acids) for each of the 239 ancestral (internal) nodes of the monocot tree. The average posterior probability (PP) for the reconstruction of all 106,116 residue positions in these sequences is 99.9%, and only 16 of these predictions have a PP  $< 80\%$ . The reconstructed sequences were used to infer 3D models of each of the ancestral octomers ( $L_8$ ) by homology, with high confidence. The stability effect of ancestral mutations was then estimated, using FoldX to make mutations in the homology model of the appropriate ancestral octomer.

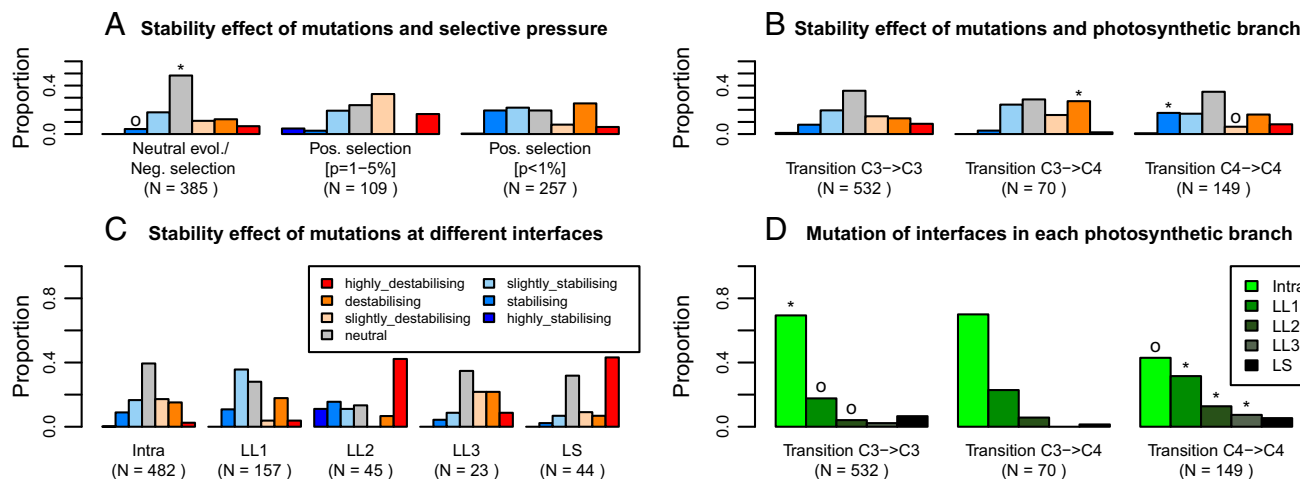
**Global analysis of the stability impact of ancestral mutations.** The distribution of the  $\Delta\Delta G_{\text{fold}}$  values of all possible mutations of *O. sativa* RubisCO (Fig. 3) is unimodal and strongly skewed toward positive values, and most possible mutations would be destabilizing. In contrast, the global distribution of  $\Delta\Delta G_{\text{fold}}$  values of the ancestral mutations follows a bimodal distribution with a high peak near zero and a smaller peak at  $+0.88$  kcal/mol (Fig. 3). Ancestral mutations are rarely strongly stabilizing or destabilizing (of the 751 in total, 6 are lower than  $-1.84$  kcal/mol and 58 are higher than  $+1.84$  kcal/mol). The vast majority of ancestral mutations (91.5%) are rather evenly distributed about zero in the  $-1.84$  to  $+1.84$  kcal/mol range, consistent with the hypothesis that maintenance of the stability of the protein is a strong constraint on evolution.

**Stability effects and selective pressures.** Among the sites that underwent mutation according to the ancestral reconstruction, two groups can be distinguished: those sites evolving under neutral evolution or negative selection and those sites under positive selection between  $C_3$  and  $C_4$  forms. Previous analyses have identified sets of 1, 2, 3, 7, 11, or 12 positively selected sites with discrepancies and overlap between the sets (Table S1). The 18 sites identified here encompass nearly all of those previously identified and 3 new sites. The sensitivity of the current analysis resolves many earlier discrepancies (Table S1) (24–26). Ancestral mutations were classified according to their evolutionary pressures (Fig. 4A), as defined by the TDG09 algorithm (27). Independently from the distinction between types of selection, mutations were also classified into three groups following the photosynthetic types of their ancestor and descendant as  $C_3 \rightarrow C_3$ ,  $C_3 \rightarrow C_4$ , and  $C_4 \rightarrow C_4$  (the change  $C_4 \rightarrow C_3$  has not been seen and detailed comparative analyses show that, if it has occurred, it must be very rare) (28).

On  $C_3 \rightarrow C_3$  branches, the distribution of stability effects follows a normal distribution, with a peak of stability-neutral mutations (Fig. 4B, Left) that preserve the global stability of the structure. In contrast, the  $C_3 \rightarrow C_4$  branches present significantly more destabilizing mutations (permutation test,  $P = 0.0080$ ; Fig. 4B, Center), which correspond to the second peak ( $+0.88$  kcal/mol) in the global distribution (Fig. 3). This tendency for destabilizing mutations to occur at the  $C_3 \rightarrow C_4$  transition is also apparent in a timeline of cumulative mutational stability changes in the ancestral sequences (Fig. 5). In  $C_4 \rightarrow C_4$  branches, a large fraction of destabilizing mutations is still observed, but there is a significantly greater proportion of mutations with a stabilizing effect compared with other branches ( $P < 0.0001$ ; Fig. 4B, Right). The timeline also shows that there is a large proportion of stabilizing mutations immediately following the  $C_3 \rightarrow C_4$  transition (Fig. 5A) and that the preponderance of stabilizing over destabilizing mutations means that the loss of stability at the transition is largely recovered within the subsequent three branches (Fig. 5B). Furthermore, considering the cumulative



**Fig. 3.** Distribution of stability effects of possible mutations and those occurring during evolution. The distribution of stability changes arising from mutations observed in the evolutionary history of the reconstructed ancestral sequences (solid line) stands in contrast to that of all possible simulated mutations (dashed line). Both distributions have their largest peak close to a  $\Delta\Delta G$  of zero. The observed mutations have an excess of slightly stabilizing observed mutations and also a distinct peak of slightly destabilizing and destabilizing values centered at  $+0.88$  kcal/mol. The majority of possible mutations are highly destabilizing and rarely occur during evolution. The probability distributions shown here are obtained by kernel smoothing of the original data (Fig. S1).



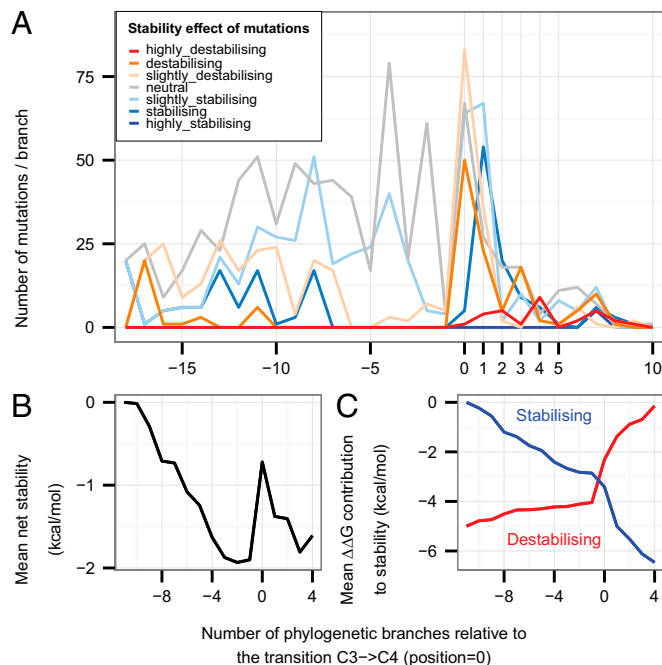
**Fig. 4.** Stability effect and location of ancestral mutations. The 751 mutations occurring during evolution are separated in **A** by their selection constraints: negative selection or neutral evolution ( $P > 0.05$  from TDG09 after false discovery rate correction), positive selection ( $0.01 < P < 0.05$ ), and strong evidence of positive selection ( $P < 0.01$ ) and binned according to their stability effect. **(B)** Mutations are separated into their branch type ( $C_3 \rightarrow C_3$ ,  $C_3 \rightarrow C_4$ , or  $C_4 \rightarrow C_4$ ) and binned by their stability effect. **(C)** Mutations are classified following the subunit interface definitions in ref. 29 and Fig. S2: Intra are in contact only with other residues of the same large subunit, LL1 residues are in contact with the other large subunit of the same dimer (e.g., the  $L_A L_B$  interface), LL2 and LL3 contact a large subunit of another dimer (e.g.,  $L_B L_C$  and  $L_B L_D$ , respectively), and LS are all residues in contact only with any of the small subunits. **(D)** Mutations are separated into their branch type and binned into their contact interfaces. Categories are highlighted by an \* when enriched or an "o" when depleted.

contributions to stability of all stabilizing mutations and all destabilizing mutations separately shows that stabilization due to all stabilizing mutations is accumulated more quickly in the branch following the  $C_3 \rightarrow C_4$  transition than at any other time (Fig. 5C).

**Stability effects and location on the 3D structure.** Ancestral mutations were grouped according to their position in the 3D structure of the hexadecamer (Fig. 4C and D) following the interface definitions in ref. 29. The stability effects of mutations within the core of the large subunit or the LL1 interfaces within dimers (e.g.,  $L_A L_B$ ) follow an approximately normal distribution. In contrast, although small in number, mutations of residues at the LL2 (e.g.,  $L_A L_H$ ) interface between dimers and at the LS interface between large and small subunits have some tendency to be highly destabilizing ( $P = 0.0318$  and  $P = 0.0053$ , respectively). The proportion of mutations at interfaces between large subunits is significantly greater in the  $C_4 \rightarrow C_4$  branches ( $P < 0.0002$ ), suggesting that the modification of subunit interactions is important for  $C_4$  optimization (Fig. 4D).

**Positively Selected Sites in the Transition to  $C_4$ .** At the  $C_3 \rightarrow C_4$  transitions, three positively selected mutations with a destabilizing effect are especially frequent: A328S, A281S, and L270I (Table S1). The A328S mutation and a positively selected, but less frequent, V326I mutation lie on either side of H327, which coordinates the P5 phosphate of the substrate in the closed state of the enzyme. Furthermore, these two residues are at the base of the active site loop (loop 6 in residues 328–337) that carries the catalytic lysine K334 and undergoes a disorder-order transition on the binding of both substrates. The replacement of hydrophobic A328 in the  $C_3$  form with a polar serine in  $C_4$  forms is destabilizing as it disrupts the packing of the base of loop 6 against  $\alpha$ -helix 6 (running from residues 338–350). This destabilization could directly alter the catalytic parameters by allowing more flexibility in loop 6, thus affecting the opening and closing of the active site (16). Extensive studies of this loop region in algal and cyanobacterial RubisCOs have shown that catalytic parameters are sensitive to its modification even if the mutated residues have no direct interaction with substrates (30). L270I is located directly beneath H298, which interacts with the P5 phosphate in the preactivated state. Replacement of V326 and L270 will also lead to packing changes that could alter the spatial disposition of the phosphate-binding histidines. Site 281 is

in the core of the C-terminal domain, and its potential to affect activity is not obvious. However, A281 packs against S321 and G322 at the end of the strand, which leads to loop 6, and destabilization



**Fig. 5.** Changes in stability through evolution. **(A)** Frequency of mutations in each category of stability against their evolutionary branch positions relative to the  $C_3 \rightarrow C_4$  transition. There is a long period in which slightly stabilizing mutations are accumulated before the transition in which a substantial number of destabilizing and slightly destabilizing mutations occur. In the branch following the transition, there is a peak of apparently compensatory stabilizing or slightly stabilizing mutations. Stability categories as in Fig. 2. **(B)** Cumulative mean net change in stability in the neighborhood of the  $C_3 \rightarrow C_4$  transition. **(C)** The corresponding cumulative mean contributions to stability of all stabilizing and all destabilizing mutations (the latter is offset by  $-5$  kcal/mol to aid comparison).

Several of the positively selected mutations found in  $C_3 \rightarrow C_4$  branches are also present in  $C_4 \rightarrow C_4$  branches, (i.e., V101I, L270I, M309I, and A328S). Additionally, three mutations on  $\alpha$ -helix 8, the final element of secondary structure of the N-terminal domain of the large subunit, are positively selected in this type of branch: P142A/T, T143A (also strongly selected in  $C_3 \rightarrow C_4$  branches), and S145A. This helix forms the symmetric interface between the N-terminal domains of large subunits on neighboring dimers (at the LL2 interfaces, e.g.,  $L_A L_H$ ). At each interface, the threonine and proline from each helix are intercalated (Fig. S3). Structural superposition of the open and closed forms of rice RubisCO suggests that an asymmetric movement of this helix between open and closed states of the upper active site, such as might occur on ligand binding or product release, will be transmitted to the neighboring active site at its lower left, potentially leading to a preference for the lower site to be closed while the top is open and vice versa.

**Diversification of RubisCO on an Island of Stability.** Throughout their evolutionary histories, RubisCO genes have faced significant changes, both internal and external to the organism, which have altered the physiologically optimal properties of RubisCO and thus the selective pressures on its evolution (10). In our example of rice RubisCO, residues at nearly all sites contribute favorably to stability, and most putative mutations would lead to destabilization (Fig. 24). The change in stability that RubisCO can withstand without dysfunction has yet to be established experimentally, but the computed stability effects of mutations that have become fixed in some species are largely confined to a narrow range near zero (Fig. 3). This small amplitude of the effects of mutation observed in nature suggests that RubisCO evolves within a small island of stability (3, 5).

Destabilizing mutations are more frequently fixed in  $C_4$  lineages. In those evolutionary branches that undergo a functional change ( $C_3 \rightarrow C_4$ ), adaptation is preceded by a long mutational sequence in which neutral to slightly stabilizing capacitive mutations dominate, i.e., which create the capacity for the protein to tolerate the destabilization required for new function (Fig. 5B). A variety of often destabilizing mutations occurs precisely at the transition to  $C_4$ , and these are immediately followed by compensatory stabilizing mutations (Fig. 5 and Fig. S4).

Except for cases in which folding is coupled to substrate binding, there is no a priori expectation of a direct physical connection between stability and activity. That similar tradeoffs between activity and stability are consistently found in both directed and natural evolution argues that an indirect connection necessarily arises from the tension between selection for optimal stability and selection for activity from a shared pool of possible mutations.

**Modulation of Conformational Change Appears to Be Key to the Adaptation of RubisCO.** Adaptive mutations occur in several distinct parts of the RubisCO structure. None are in direct contact with the substrates; however, a small number of second shell mutations (i.e., residues in contact with active site residues) are strongly positively selected. These mutations tend to be destabilizing and, on the basis of structural context and earlier mutational studies of algal RubisCOs, are inferred to modify the active site loop dynamics or position of residues at the P5 and O<sub>2</sub>/CO<sub>2</sub> binding sites. Whereas adaptive mutations 10–20 Å from active sites have occasionally been identified in other enzymes (31), in RubisCO, these form the majority of positively selected sites that distinguish C<sub>3</sub> and C<sub>4</sub> species. Experiments with RubisCO from the green alga *Chlamydomonas reinhardtii* previously implicated the interfaces between large and small subunits in the modulation of catalytic rates (32). The analysis here increases the number of known functionally significant intersubunit sites (Table S2) and demonstrates a link with the C<sub>3</sub>-C<sub>4</sub> transitions in flowering plants. Those mutations near the dimer or N- and C-terminal domain interfaces within each large subunit likely affect the substantial relative movements of the domains on substrate binding. Although one of these residue changes (M309I) has previously been shown to switch the enzyme to C<sub>4</sub>-like properties in plants (21), it is clear that this change is not essential, and there are other mutational routes to equivalent functional changes.

**Altered Cooperativity May Have an Adaptive Role in Some Species.** Negative cooperativity has been reported for the binding of the transition-state analog 2-carboxyarabinitol biphosphate to the active site of the C<sub>3</sub> RubisCO from spinach (33). Kinetic data fit a model of rapid binding to one half of the active sites accompanied by the slower binding to the remainder (34). Although it has proven difficult to generalize these observations to other species (possibly because of the stringent demands for pure and active protein in such experiments and because weak negative cooperativity is also intrinsically difficult to unambiguously identify in standard turnover kinetics), they naturally led to a postulated enzymatic mechanism whereby binding of substrates to one site of each dimer reduces binding at the other (34). Crystallographic studies have not been able to directly address this issue as they produce symmetric structures, either apo or fully saturated (16). The observation of positive selection on mutations in the interface between the N-terminal domains of neighboring dimers suggests a different mechanism of cooperativity. Comparison of hybrid structures of apo and holo forms of RubisCO suggests that conformational changes at an active site in the ring of active sites at the top of the oligomer are coupled to the lower site in the dimer to its left. The mutations occurring during the C<sub>3</sub> to C<sub>4</sub> transitions diminish this coupling and would relieve any negative cooperativity between the upper and lower sites, thus enhancing turnover. The identified positive selection suggests that these mutations play a role in the adaptation of some C<sub>4</sub> species. Consequently, these mutations and the possibility of a role for cooperativity in RubisCO warrant renewed experimental investigation.



**Conclusions.** The mutational landscape of RubisCO is strongly constrained by the need to maintain overall stability. This constraint limits the adaptation of RubisCO to novel environmental contexts to those amino acid changes that can modify the catalytic efficiency without dramatic effect on the overall folding stability. Following the repeated origins of  $C_4$  photosynthesis in flowering plants, a number of amino acid mutations of RubisCO were preferentially kept by natural selection. These mutations include changes to residues that might modify the geometry of the active site, as well as a substantial number of sites at the interface between domains and subunits, which probably alter the properties of the enzyme via modification of the dynamics of conformational change or alteration of cooperativity between catalytic subunits. It is clear that a substantial proportion of the mutations necessary for  $C_4$  adaptation are themselves destabilizing. Evolution accommodates such destabilizing functional adaptations thanks to the previous accumulation of stabilizing capacitive mutations and by subsequently fixing stabilizing compensating mutations.

## Methods

The multiple sequence alignment of genes for RubisCO large subunit (*rbcl*) and its associated phylogenetic tree are from Christin et al. (24). The highest-resolution (1.35 Å) structure of RubisCO currently available, from the  $C_3$  grass *Oryza sativa* (35), was used as the basis for structural analyses. The complete biological unit ( $L_6S_6$ ) was directly downloaded from the PDBEISA website (36).

The Protein Data Bank (PDB) structure file for the large subunit contains coordinates for residues 11–475 (465 residues). This structure was used as a template for the homology modeling of 3D octomeric structures ( $L_8$ ) of each ancestral *rbcl* sequence. The modeling was done with Modeler 9.9 (37). For each sequence, 100 models were built, and the model with the lowest energy (based on its discrete optimized protein energy score) was used in further analyses. Using FoldX 3b5.1 (38), the energies for the WT ( $\Delta G_{\text{fold,wt}}$ ) and mutant ( $\Delta G_{\text{fold,mu}}$ ) protein were computed to give the stability change  $\Delta\Delta G_{\text{fold}} = \Delta G_{\text{fold,mu}} - \Delta G_{\text{fold,wt}}$ . The SD in FoldX is 0.46 kcal/mol (38), and we used this value to bin the  $\Delta\Delta G_{\text{fold}}$  values into seven categories. Additional FoldX restraints were applied to the conserved active site to avoid the potential for artifacts arising from unparameterised ligands. The inference of ancestral sequences was performed under maximum likelihood as implemented in CodeML (39). Sites under positive selection between  $C_3$  and  $C_4$  forms were identified by the TDG09 algorithm (27), which performs a likelihood ratio test to assess if the evolutionary rate at a particular position is similar or different between  $C_3$  and  $C_4$  lineages. The  $\Delta\Delta G_{\text{fold}}$  due to each mutation on each branch was then mapped onto the phylogenetic tree (Fig. S4). Detailed methods are given in SI Text.

**ACKNOWLEDGMENTS.** This study benefited from use of the University College London (UCL) Legion High-Performance Computing Facility (Legion@UCL). R.A.S. acknowledges funding from the Fondation du 450ème Anniversaire de l'Université de Lausanne and Swiss National Science Foundation Grants 132476 and 136477. P.-A.C. is funded by Marie Curie International Outgoing Fellowship 252568.

- Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10(12):866–876.
- DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat Rev Genet* 6(9):678–687.
- Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46(1):105–109.
- Tokuriki N, Tawfik DS (2009) Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 19(5):596–604.
- Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Comput Biol* 4(2):e1000002.
- Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11(8):572–582.
- Wang X, Minasov G, Shoichet BK (2002) Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol* 320(1):85–95.
- Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103(15):5869–5874.
- Bloom JD, Arnold FH (2009) In the light of directed evolution: Pathways of adaptive protein evolution. *Proc Natl Acad Sci USA* 106(Suppl 1):9995–10000.
- Tcherkez GG, Farquhar GD, Andrews TJ (2006) Despite slow catalysis and confused substrate specificity, all ribulose biphosphate carboxylases may be nearly perfectly optimized. *Proc Natl Acad Sci USA* 103(19):7246–7251.
- Lorimer GH, Andrews TJ (1973) Plant photorespiration—An inevitable consequence of the existence of atmospheric oxygen. *Nature* 243(5406):359–360.
- Sage RF, Sage TL, Kocacinar F (2012) Photorespiration and the evolution of  $C_4$  photosynthesis. *Annu Rev Plant Biol* 63:19–47.
- Sage RF, Christin PA, Edwards EJ (2011) The  $C(4)$  plant lineages of planet Earth. *J Exp Bot* 62(9):3155–3169.
- Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers  $C_4$  origins. *New Phytol* 193(2):304–312.
- Savir Y, Noor E, Milo R, Tlustý T (2010) Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci USA* 107(8):3475–3480.
- Andersson I, Backlund A (2008) Structure and function of Rubisco. *Plant Physiol Biochem* 46(3):275–291.
- Young JN, Rickaby RE, Kapralov MV, Filatov DA (2012) Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. *Philos Trans R Soc Lond B Biol Sci* 367(1588):483–492.
- Sage RF (2002) Variation in the  $k(\text{cat})$  of Rubisco in  $C(3)$  and  $C(4)$  plants and some implications for photosynthetic performance at high and low temperature. *J Exp Bot* 53(369):609–620.
- Hudson GS, et al. (1990) Comparisons of *rbcl* genes for the large subunit of ribulose-biphosphate carboxylase from closely related  $C_3$  and  $C_4$  plant species. *J Biol Chem* 265(2):808–814.
- Kapralov MV, Kubien DS, Andersson I, Filatov DA (2011) Changes in Rubisco kinetics during the evolution of  $C_4$  photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Mol Biol Evol* 28(4):1491–1503.
- Whitney SM, et al. (2011) Isoleucine 309 acts as a  $C_4$  catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in Flaveria. *Proc Natl Acad Sci USA* 108(35):14688–14693.
- Dessailly BH, Lensink MF, Wodak SJ (2007) Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics* 8:141.
- Beadle BM, Shoichet BK (2002) Structural bases of stability-function tradeoffs in enzymes. *J Mol Biol* 321(2):285–296.
- Christin PA, et al. (2008) Evolutionary switch and genetic convergence on *rbcl* following the evolution of  $C_4$  photosynthesis. *Mol Biol Evol* 25(11):2361–2368.
- Wang M, Kapralov MV, Anisimova M (2011) Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. *BMC Evol Biol* 11:266.
- Kapralov MV, Filatov DA (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol* 7:73.
- Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA (2009) Identifying changes in selective constraints: Host shifts in influenza. *PLOS Comput Biol* 5(11):e1000564.
- Christin PA, Freckleton RP, Osborne CP (2010) Can phylogenetics identify  $C(4)$  origins and reversals? *Trends Ecol Evol* 25(7):403–409.
- van Lun M, van der Spoel D, Andersson I (2011) Subunit interface dynamics in hexadecameric rubisco. *J Mol Biol* 411(5):1083–1098.
- Parry MA, Andralojc PJ, Mitchell RA, Madgwick PJ, Keys AJ (2003) Manipulation of Rubisco: The amount, activity, function and regulation. *J Exp Bot* 54(386):1321–1333.
- Thomas VL, McReynolds AC, Shoichet BK (2010) Structural bases for stability-function tradeoffs in antibiotic resistance. *J Mol Biol* 396(1):47–59.
- Spreitzer RJ, Peddi SR, Satagopan S (2005) Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. *Proc Natl Acad Sci USA* 102(47):17225–17230.
- Johal S, Partridge BE, Chollet R (1985) Structural characterization and the determination of negative cooperativity in the tight binding of 2-carboxyarabinitol biphosphate to higher plant ribulose biphosphate carboxylase. *J Biol Chem* 260(17):9894–9904.
- Zhu G, Jensen RG (1990) Status of the substrate binding sites of ribulose biphosphate carboxylase as determined with 2-C-carboxyarabinitol 1,5-bisphosphate. *Plant Physiol* 93(1):244–249.
- Matsumura H, et al. (2012) Crystal structure of rice Rubisco and implications for activation induced by positive effectors NADPH and 6-phosphogluconate. *J Mol Biol* 422(1):75–86.
- Krisinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372(3):774–797.
- Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234(3):779–815.
- Schymkowitz J, et al. (2005) The FoldX web server: An online force field. *Nucleic Acids Res* 33(Web Server issue):W382–8.
- Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.

# Supporting Information

Studer et al. 10.1073/pnas.1310811111

## SI Text

**Description of the Biological Unit of Ribulose-1,5-Bisphosphate Carboxylase.** In land plants, the ribulose-1,5-bisphosphate carboxylase (RubisCO) biological unit is an oligomer of 16 subunits. Two large (L) subunits head-to-tail form a  $L_2$  dimer. Four  $L_2$  dimers form the main octomeric ring ( $L_8$ ). There are also four small subunits on the top and four small subunits on the bottom that together with the large units form the hexadecamer  $L_8S_8$  complex, which is the biological unit (1, 2). The catalytic activity of RubisCO occurs in the large subunit, and the main component of this large subunit is the C-terminal  $\alpha/\beta$  barrel domain (positions 169–432), in which the catalytic residues lie (3). Various residues are important for catalysis. Residues 201, 203, and 204 are charged residues and bind a magnesium atom in the catalytic site. Residues 175 and 295 play the role of proton acceptors. Residues 123, 173, 177, 295, 327, and 379 form the binding pocket. Site 334 is a transition state stabilizer. The sixth loop (328–337) is an important component as it participates in the opening and closure of the active site (2). We use the nomenclature of ref. 4 for labeling the different interfaces LL1, LL2, and LL3 (Fig. S2).

**Calculation of Stability Effects of Mutation with FoldX.** All stability analyses were performed with FoldX 3b5.1 (5, 6), which is one of the best stability predictors and easily implementable in a pipeline (7, 8). No current stability prediction method is very accurate in predicting the effect of individual mutations [evaluations of FoldX have reported a Matthews correlation coefficient ( $r$ ) of 0.81 (5) and 0.59 (8) for individual mutations dependent on the protein test set]. However, a number of structure-based methods do predict the trend in average values for binned data very well. FoldX achieved the highest correlation ( $r = 0.96$ ) for binned data in a recent evaluation (8). The analysis reported here relies on aggregating results of many mutations in different evolutionary or structural contexts, and it is consequently expected that observed trends in stability will be robust. We used the reported accuracy of FoldX of 0.46 kcal/mol (i.e., the SD of the difference between  $\Delta\Delta G$ s calculated by FoldX and the experimental values) (5) to bin the  $\Delta\Delta G$  values into seven categories: (i) highly stabilizing ( $\Delta\Delta G < -1.84$  kcal/mol); (ii) stabilizing ( $-1.84$  kcal/mol  $\leq \Delta\Delta G < -0.92$  kcal/mol); (iii) slightly stabilizing ( $-0.92$  kcal/mol  $\leq \Delta\Delta G < -0.46$  kcal/mol); (iv) neutral ( $-0.46$  kcal/mol  $< \Delta\Delta G \leq +0.46$  kcal/mol); (v) slightly destabilizing ( $+0.46$  kcal/mol  $< \Delta\Delta G \leq +0.92$  kcal/mol); (vi) destabilizing ( $+0.92$  kcal/mol  $< \Delta\Delta G \leq +1.84$  kcal/mol); and (vii) highly destabilizing ( $\Delta\Delta G > +1.84$  kcal/mol).

Before the analysis, all of the chains of the *Oryza sativa* RubisCO structure [Protein Data Bank (PDB) code: 1WDD] were distinctly renamed, and all heteroatoms in the structure that are not well parameterized in FoldX [2-Cabp (CAP), glycerol (GOL), and  $Mg^{2+}$  (MG)] were removed. Although  $Mg^{2+}$  is important to the overall structure of RubisCO, because  $Mg^{2+}$ 's short-range interactions are not well parameterized in FoldX, we mimic the conformational effects of  $Mg^{2+}$  by constraints that maintain the binding site conformation very close to that observed experimentally. These constraints avoid potential artifacts due to an inaccurate  $Mg^{2+}$  model and are practicable in the case of RubisCO as all residues surrounding the  $Mg^{2+}$  atom are 100% conserved. None participate in the differences between  $C_3$  and  $C_4$  forms of RubisCO. The closest residue under selection between  $C_3$  and  $C_4$  is 328, but the stability effect of the mutation A328S is very similar whether the magnesium is present or not.

The binding site restraints are, of course, the result of compromise and may themselves be a source of some inaccuracy of predictions for some residues as local structure relaxation of second shell residues bordering the binding site could potentially be impeded. To perform the stability analyses, we use the standard pipeline for FoldX, which uses two commands: Repair and Build-Model. The Repair command was used to optimize the structure by removing any steric clashes. This optimization improves the global stability score of the RubisCO  $L_8S_8$  complex, reducing the free energy ( $\Delta G$ ) from +105.72 kcal/mol (before the repair step) to -1,506.42 kcal/mol (after the repair step). For the analysis of all possible mutations, each residue of the *O. sativa* RubisCO structure was mutated using the command BuildModel. FoldX operates in two steps, repeated five times (option <number-OfRuns>5) to ensure that the minimum energy conformations of even large residues that possess many rotamers is identified: Step 1, the residue of interest and its neighbors are mutated to themselves and conformationally relaxed to remove any local clash and the stability of the WT ( $\Delta G_{wt}$ , in kcal/mol) is obtained by the FoldX energy function; Step 2, the residue is mutated to all 19 other amino acids and all other neighbor side chains to themselves. The stability of the mutant ( $\Delta G_{mut}$ , in kcal/mol) is calculated.

Each mutation is made in all large subunits (LSUs) simultaneously, and the reported stability effect of a mutation  $\Delta\Delta G = \Delta G_{mut} - \Delta G_{wt}$  is normalized for the number of large subunits (by dividing the calculated change by eight). The results of these calculations are presented in Fig. 2 and Table S3.

Whereas calculations were actually carried out for 463 residues of the *O. sativa* RubisCO structure (residues 12–475, which are present in all rbcL subunits less the essential modified residue Lcx201), for the sake of simplifying the presentation (and not affecting any functionally significant residue), the mutational data are presented in text, figures, and tables only for those residues that are common to both reconstructed sequences (i.e., those positions that can be continuously aligned) and structural models, specifically, Gly12 through Ala456 (less the essential Lcx201), i.e., in the case of this stability analysis for 8,436 simulated mutations (19 amino acids  $\times$  444 residues).

**Ancestral Sequence Reconstruction.** Recent analyses have successfully used ancestral sequence reconstruction of extinct proteins to reproduce their evolutionary history in vitro and/or in silico (9, 10). We used the collection of extant large subunit RubisCO protein sequences and the phylogenetic tree for monocot species from the study of Christin et al. (11). The alignment contains 240 sequences [137 of  $C_3$ , 101 of  $C_4$ , and 2 of crassulacean acid metabolism (CAM) plants], all from monocots, one of the major groups of flowering plants. Ancestral sequence reconstruction of extinct RubisCO proteins was performed under maximum likelihood as implemented in CodeML (PAML package release 4.4d) (12).

**Stability Effects of Ancestral Mutations.** Homology models of the ancestral octomers were constructed for each branch point of the phylogenetic tree. The stability effect of ancestral mutations occurring in each branch of the tree was estimated using FoldX, as described above, to make the mutations in the appropriate ancestral octomer structure. Consequently, each mutation in a phylogenetic branch is evaluated in the context of a structural model containing all of the accumulated mutations up to the branch point,



and many of the possible interactions between mutations that accumulate during evolution are accounted for.

However, in those cases where multiple mutations occur within a branch, the reconstruction procedure cannot tell us the order in which they occur. Consequently, there is a possibility that if mutations within a branch interact, their historical order of occurrence could alter their  $\Delta\Delta G$  contribution. It is difficult to test all possible orders (and present the results of that analysis); consequently, in the main results, we assume that there is no interaction between mutations within a branch.

To test the reasonableness of this assumption, we compared, for all 172 branches with multiple mutations, the sum of the  $\Delta\Delta G$  of individual mutations to the  $\Delta\Delta G$  of all mutations within a branch made simultaneously. The differences between these sums of individual and simultaneous mutations are shown in Fig. S5. Because many pairs or triples of ancestral mutations are spatially separated in the structure, the vast majority of groups of mutations in a branch are approximately additive. In only a few cases (15/172) is this difference greater than the FoldX SD of 0.46 kcal/mol, and only one of these lies on a  $C_3 \rightarrow C_4$  branch. Consequently, it seems reasonable to conclude that any effect of nonadditivity within branches will not significantly impact on the analysis.

The distributions of the stability effects computed with FoldX for all possible mutations and for the ancestral mutations are shown in Fig. 3 and Fig. S1.

**Identification of Sites Under Positive Selection.** The fixation rate of mutations in the genome depends on selective forces and genetic drift. This rate will be accelerated under positive selection (i.e., to promote a new function) or decelerated under negative selection (i.e., when a mutation has a negative impact on the organismal fitness). Positions under selective constraints are likely to be important in explaining functional changes between subgroups, such as between  $C_3$  and  $C_4$  plants. The TDG09 algorithm (13) was used to detect sites experiencing changes in selective constraints between  $C_3$  and  $C_4$  forms. It requires as input a multiple alignment of homologous sequences, a well-defined phylogenetic tree, and a grouping of these sequences into two categories (in our cases,  $C_3$  and  $C_4$ ). The method estimates parameters for each site using two assumptions. The first assumption is that evolutionary rates can vary between positions, but these evolutionary rates are constant within a given position across species (constant evolutionary rate over time), i.e., the homogeneous model of substitutions (model 2 as described in ref. 13). The second assumption is that evolutionary rates can vary at a given position depending on the species group (shift in evolutionary rate over time), i.e., the nonhomogeneous model of substitutions (model 3). For each site, TDG09 computes the likelihood of both model 2 and model 3 and compares them by a likelihood ratio test (LRT), which will give a  $P$  value per site. A false discovery rate (FDR) correction is applied. Before the analysis, the lengths of all branches were estimated with CodeML (PAML package release 4.4d) (12) under the Whelan and Goldman model (14). The two CAM sequences (*Ananas* and *Tillandsia*) were omitted for the purpose of this analysis, as they cannot be clearly assigned to either the  $C_3$  or the  $C_4$  group and may be under distinct selective pressures. We used relaxed and stringent FDR thresholds (5% and 1%, respectively) to select sites under functional divergence from the TDG09 output. We identified 18 sites under positive selection at the 5% FDR level, with a subset of 12 sites at the 1% FDR level (Table S1).

In their previous study, using the same dataset, Christin et al. (11) identified 14 sites under codon substitution models (site and branch-site). Codon substitution models use nucleotide sequences to compute the dN/dS ratio of non-synonymous (dN) substitutions over synonymous (dS) substitutions to infer positive selection. Our analysis of this dataset, analyzed with the TDG09

algorithm (using evolutionary models based on amino acid sequences), identified 18 sites under positive selection between  $C_3$  and  $C_4$  forms (Table S1). Nine sites had been detected under the codon substitution site model M2a in Christin et al. (11). TDG09 detected seven of them with the exception of positions 91 and 265. The branch-site model identified five sites with purifying selection in  $C_3$  forms and positive selection in  $C_4$  (101, 258, 270, 281, and 309). TDG09 identified all of them at the 1% FDR level. Three sites evolved under neutral evolution in  $C_3$  and under positive selection in  $C_4$  (142, 145, and 328). TDG09 identified two of the sites at 1% (145 and 328), whereas position 142 is only detected at the 5% FDR level. In addition, TDG09 identified three new sites at 1% levels (143, 225, and 262) and three at 5% (221, 282, and 326), which were not detected by the codon substitution models used by Christin et al. (11).

Some of the sites under positive selection have been detected in previous studies conducted in conifers (15) and flowering plants (16, 17), such as in monocots (11, 18) and eudicots (19, 20) (Table S1).

It is interesting that the codon substitution methods correlate well because codon substitution models can be affected by saturation at the dS level (i.e., many synonymous substitutions will be seen only as one substitution), and results can be less reliable as the evolutionary time increases, especially in highly divergent gene families. Evolutionary methods based on amino acids, such as the TDG09 algorithm, can help to override this problem.

**Identification of Coevolutionary Sites Under Positive Selection.** Coevolutionary information was detected with EVfold (21) and for several datasets of *rbcl* sequences (Table S2).

**Extended Description of Sites Under Positive Selection.** In general, more mutations are observed on average on  $C_3 \rightarrow C_4$  branches (3.7 per branch) than in  $C_3 \rightarrow C_3$  branches (2.8) and  $C_4 \rightarrow C_4$  branches (2.0).

**$C_3 \rightarrow C_3$  branches.** On  $C_3 \rightarrow C_3$  branches, several sites are under positive selection. The mutations I225L, A228S, and V262A are found at or near the interface between the C-terminal domain of the large subunit and the small subunit to its right (Fig. S3), and mutations of I251L and A/V255I/T at the base of the  $\alpha/\beta$ -barrel are also common. These mutations have a mixed effect on predicted stability, but are usually destabilizing (Table S1).

Despite their being far from the active site, mutations of residues in these regions of the structure are positively selected in all branches and have been shown to affect the catalytic properties of RubisCO from the green alga *Chlamydomonas reinhardtii*. Genetic screening of *C. reinhardtii* has shown that A222T or V262L mutations restore thermal stability and  $CO_2$  specificity to a temperature-sensitive mutant (22). However, in *C. reinhardtii*, the mutation R258K (which is positively selected in  $C_4 \rightarrow C_4$  branches) is at the interface of large ( $LL_3$ ) and small ( $LS_1$ ) subunits and is associated with 23% and 33% increases in maximum reaction rate values for oxygenation ( $V_o$ ) and carboxylation ( $V_c$ ), respectively (23). A penta-mutant V221C/V235I/C256F/K258R/I265V changes individual  $K_m$  and  $k_{cat}$  values for  $O_2$  and  $CO_2$  by factors of 2–3 while leaving specificity unchanged (24). Fig. S3B shows the close proximity of the group of positively selected residues, found at the  $L_B S_B$  interface (and to a lesser extent the  $L_B L_D$  interface) and at the base of the  $\alpha/\beta$  barrel (in a contiguous stretch of sequence from 221 to 282).

The other positively selected mutation in  $C_3 \rightarrow C_3$  branches is S328A near the active site (discussed further below). This mutation has a slightly stabilizing effect.

**$C_3 \rightarrow C_4$  branches. Mutations close to the active site.** Two mutations with a destabilizing effect are especially frequent (Table S1): L270I and A328S. The mutations A328S and the positively selected but less frequent V326I lie on either side of H327, which coordinates the P5 phosphate of the substrate in the closed state of the enzyme. Furthermore, these two residues are at the base of the active site

loop (loop-6 residues 328–337) that carries the catalytic lysine-334 and makes a disorder-order transition on the binding of both substrates. L270I is located directly beneath H298, which interacts with the P5 phosphate in the preactivated state.

The replacement of hydrophobic alanine in the C<sub>3</sub> form at 328 with a polar serine in C<sub>4</sub> forms is destabilizing as it disrupts the packing of the base of loop 6 against helix 6 (running from residues 338 to 350). Replacement of V326 and L270 will also lead to packing changes that could alter the spatial disposition of the phosphate-binding histidines. The mutation of site 328 from a hydrophobic to a polar residue could also directly alter the catalytic parameters by allowing more flexibility in the sixth loop, altering the kinetics of opening and closing of the active site. Extensive studies of this loop region in algal and cyanobacterial RubisCOs show that catalytic parameters are sensitive to its modification even if the mutated residues have no direct interaction with substrates (25). In particular, the mutations V331A and L326I in the RubisCO in *C. reinhardtii*, which also modify the loop 6/helix 6 interface, are both known to decrease CO<sub>2</sub> specificity (26, 27).

**Mutations that have an impact on the active site.** Under strong positive selection in C<sub>3</sub>→C<sub>4</sub> and C<sub>4</sub>→C<sub>4</sub> is the mutation V101I. This residue is always valine in C<sub>3</sub> monocots, and the addition of one carbon to the side chain could shift the second  $\alpha$ -helix in the N-terminal domain toward the active site. This movement is significant as directly on the opposite side of this helix is glutamate-60, which forms a salt bridge with the catalytic K332 in the closed activated state of the enzyme. Consequently, any movement of the  $\alpha$ -helix could affect the geometry of the CO<sub>2</sub>-bound and transition states of the reaction.

**Mutations at the interface between subunits.** The mutation M309I is under strong positive selection as also identified in some previous studies (11, 19). This mutation has a neutral effect on stability but lies at the interface of the two large subunits that make up the functional dimers and also close to the junction between N- and C-terminal domains. This mutation has been demonstrated to act as a catalytic switch between C<sub>3</sub>-like and C<sub>4</sub>-like behaviors (i.e., decreasing specificity for CO<sub>2</sub> over O<sub>2</sub> and increasing the turnover) in *Flaveria* species and in chimeric enzymes consisting of large subunits from *Flaveria* and tobacco small subunits (28). However, sequence analysis shows that the isoleucine is only present in half of all of the monocot C<sub>4</sub> forms, suggesting that there are other mechanisms for making similar catalytic changes and/or that the switch may require a particular context in which to function. Interestingly, site 328, which neighbors the P5 binding H327, is coupled to 309 in angiosperm evolution (Table S2), suggesting a coupling of ribulose-1,5-bisphosphate (RBP) binding to conformational change of the dimer involving these residues. [There is also evidence from monocot species that site 326 is coupled to 362 near the interface of N- and C-terminal domains of the large subunit, also suggesting a coupling of binding to conformational change. However, mutations of this latter pair (I→V) are both positively selected in C<sub>3</sub>→C<sub>3</sub> branches, perhaps indicating an opposite effect on activity.]

A281S is distant from the active site, but has high penetration in the C<sub>4</sub> monocot species in our dataset (67% S) and is also prevalent in eudicot species (29). It forms part of the well-packed and rigid core of the C-terminal domain, and the substitution is predicted to be destabilizing. This substitution may result in increased flexibility or conformational changes locally, which could potentially be significant globally as the methyl group

of alanine is in direct contact with S321 and G322 at the end of the strand, which leads to loop 6 and with Q149 at the beginning of the loop that connects the N- and C-terminal domains.

**C<sub>4</sub>→C<sub>4</sub> branches.** Some mutations found in C<sub>3</sub>→C<sub>4</sub> branches (i.e., V101I, L270I, M309I, and A328S) are also present in C<sub>4</sub>→C<sub>4</sub> branches. Three mutations on  $\alpha$ -helix 8, the final element of secondary structure of the N-terminal domain of the large subunit, are positively selected in this branch: P142A/T, T143A (also strongly positively selected in the C<sub>3</sub>→C<sub>4</sub> branches), and S145A.  $\alpha$ -Helix 8 forms the symmetric interface between the N-terminal domains of large subunits on neighboring dimers. At the interface, the threonine and proline from each helix are intercalated (Fig. S3C). Structural superposition of the open (PDB code: 3AXM) and closed (PDB code: 1WDD) forms of rice RubisCO (Fig. S3C) suggests that an asymmetric movement of this helix between open and closed states of the upper active site, such as might occur on ligand binding or product release, will be transmitted to the neighboring active site at its lower left, leading to a preference for the lower site to be closed while the top is open and vice versa.

In this branch, one mutation, R258K, is positively selected at the interface of large (LL<sub>3</sub>) and small (LS<sub>1</sub>) subunits. In *C. reinhardtii*, this change is associated with 23% and 33% increases in maximum reaction rate values for V<sub>o</sub> and V<sub>c</sub>, respectively (23). Furthermore, this residue interacts with the tip of a long loop in the rbcS subunit (positions 56–58). Residue 57 of the rbcS has been found to be under weak selective pressure in the C<sub>3</sub>→C<sub>4</sub> transition in *Flaveria* species (19), and residues 56 and 58 are mutated between C<sub>3</sub> rice and C<sub>4</sub> corn, supporting a widespread role for interaction with the small subunit in this region modulating kinetics in the C<sub>4</sub> transition.

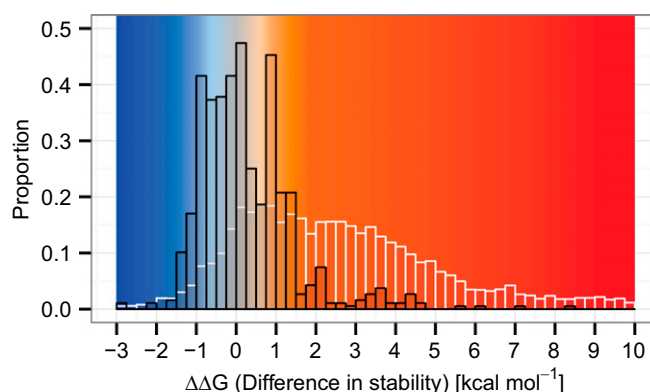
**Statistical Analysis and Visualization.** Python, Biopython (30), and R (31) were used to prepare the data and perform the computational and statistical analyses. The multiple sequence alignment was visualized with Jalview (32). The phylogenetic tree was annotated with EvolView (33). The visualizations of sites in 3D were created with Chimera (34).

To test for statistical significance of the observed variations in distributions of stability in different photosynthetic branches, under different selective forces, and in different interfaces (Fig. 4 A–C) or location (Fig. 4D) of mutations, permutation tests were carried out to investigate the null hypothesis that the distributions in each category are sampled from the same underlying distribution. For example, in comparing the distributions with respect to photosynthetic branch, we randomly permute the branch labels of all mutations 100,000 times and then count the number of times the simulated frequencies for a given stability effect are higher than the observed sample in each category. The result is a *P* value for each stability effect, which can be compared using a two-sided test, where for a single sample, a *P* < 0.05/2 = 0.025 indicates a significant enrichment for that particular stability effect and *P* > 0.975 indicates a significant depletion for that particular effect. The *P* values for all cases are given in Table S4. The permutation of the interface mutations takes into account the number of sites in each interface. When attributing significance in this study, we also conservatively corrected for the potential FDR by dividing the *P* value threshold by 3 when analyzing the number of selection pressure categories (Fig. 4A) or branches (Fig. 4B and D) or by 5 when considering interfaces (Fig. 4C).

- Andersson I, Taylor TC (2003) Structural framework for catalysis and regulation in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Arch Biochem Biophys* 414(2):130–140.
- Andersson I, Backlund A (2008) Structure and function of Rubisco. *Plant Physiol Biochem* 46(3):275–291.
- Chapman MS, et al. (1988) Tertiary structure of plant RuBisCO: Domains and their contacts. *Science* 241(4861):71–74.

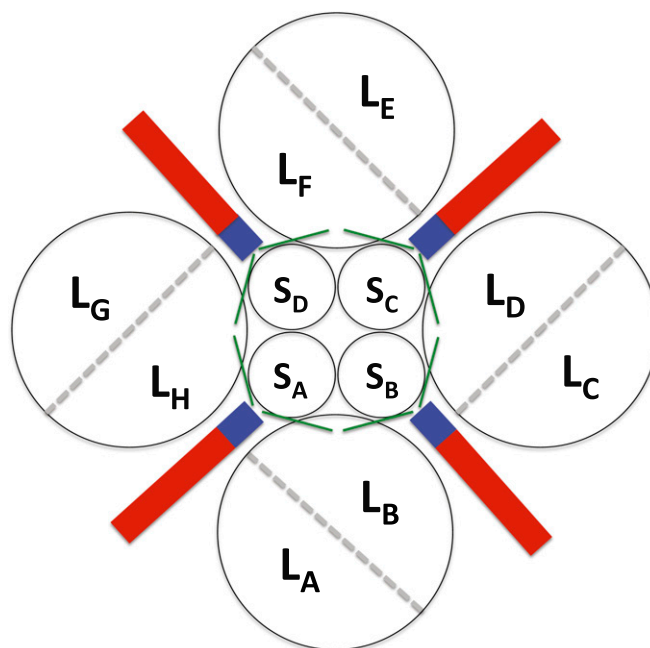
- van Lun M, van der Spoel D, Andersson I (2011) Subunit interface dynamics in hexadecameric rubisco. *J Mol Biol* 411(5):1083–1098.
- Schymkowitz J, et al. (2005) The FoldX web server: An online force field. *Nucleic Acids Res* 33(Web Server issue):W382–8.
- Schymkowitz JW, et al. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc Natl Acad Sci USA* 102(29):10147–10152.

7. Khan S, Vihinen M (2010) Performance of protein stability predictors. *Hum Mutat* 31(6):675–684.
8. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: Good on average but not in the details. *Protein Eng Des Sel* 22(9):553–560.
9. Liberles DA (2007) *Ancestral Sequence Reconstruction* (Oxford Univ Press, Oxford, UK).
10. Harms MJ, Thornton JW (2010) Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20(3):360–366.
11. Christin PA, et al. (2008) Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol Biol Evol* 25(11):2361–2368.
12. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
13. Tamuri AU, Dos Reis M, Hay AJ, Goldstein RA (2009) Identifying changes in selective constraints: Host shifts in influenza. *PLOS Comput Biol* 5(11):e1000564.
14. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18(5):691–699.
15. Sen L, et al. (2011) Molecular evolution of rbcL in three gymnosperm families: Identifying adaptive and coevolutionary patterns. *Biol Direct* 6:29.
16. Kapralov MV, Filatov DA (2007) Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol Biol* 7:73.
17. Wang M, Kapralov MV, Anisimova M (2011) Coevolution of amino acid residues in the key photosynthetic enzyme Rubisco. *BMC Evol Biol* 11:266.
18. Iida S, et al. (2009) Molecular adaptation of rbcL in the heterophyllous aquatic plant *Potamogeton*. *PLoS ONE* 4(2):e4633.
19. Kapralov MV, Kubien DS, Andersson I, Filatov DA (2011) Changes in Rubisco kinetics during the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Mol Biol Evol* 28(4):1491–1503.
20. Liu L, Zhao B, Zhang Y, Wang J (2012) Adaptive evolution of the rbcL gene in Brassicaceae. *Biochem Syst Ecol* 44(0):13–19.
21. Marks DS, et al. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6(12):e28766.
22. Hong S, Spreitzer RJ (1997) Complementing substitutions at the bottom of the barrel influence catalysis and stability of ribulose-bisphosphate carboxylase/oxygenase. *J Biol Chem* 272(17):11114–11117.
23. Du YC, Peddi SR, Spreitzer RJ (2003) Assessment of structural and functional divergence far from the large subunit active site of ribulose-1,5-bisphosphate carboxylase/oxygenase. *J Biol Chem* 278(49):49401–49405.
24. Spreitzer RJ, Peddi SR, Satagopan S (2005) Phylogenetic engineering at an interface between large and small subunits imparts land-plant kinetic properties to algal Rubisco. *Proc Natl Acad Sci USA* 102(47):17225–17230.
25. Parry MA, Andralojc PJ, Mitchell RA, Madgwick PJ, Keys AJ (2003) Manipulation of Rubisco: The amount, activity, function and regulation. *J Exp Bot* 54(386):1321–1333.
26. Chen ZX, Spreitzer RJ (1989) Chloroplast intragenic suppression enhances the low CO<sub>2</sub>/O<sub>2</sub> specificity of mutant ribulose-bisphosphate carboxylase/oxygenase. *J Biol Chem* 264(6):3051–3053.
27. Zhu G, Spreitzer RJ (1996) Directed mutagenesis of chloroplast ribulose-1,5-bisphosphate carboxylase/oxygenase. Loop 6 substitutions complement for structural stability but decrease catalytic efficiency. *J Biol Chem* 271(31):18494–18498.
28. Whitney SM, et al. (2011) Isoleucine 309 acts as a C4 catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in *Flaveria*. *Proc Natl Acad Sci USA* 108(35):14688–14693.
29. Kapralov MV, Smith JA, Filatov DA (2012) Rubisco evolution in *C. eudicot*: An analysis of *Amaranthaceae* sensu lato. *PLoS ONE* 7(12):e52974.
30. Cock PJ, et al. (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.
31. R Development Core Team (2011) *R: A Language and Environment for Statistical Computing* (R Development Core Team, Vienna).
32. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.
33. Zhang H, Gao S, Lercher MJ, Hu S, Chen WH (2012) EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Res* 40(Web Server issue):W569–72.
34. Pettersen EF, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612.

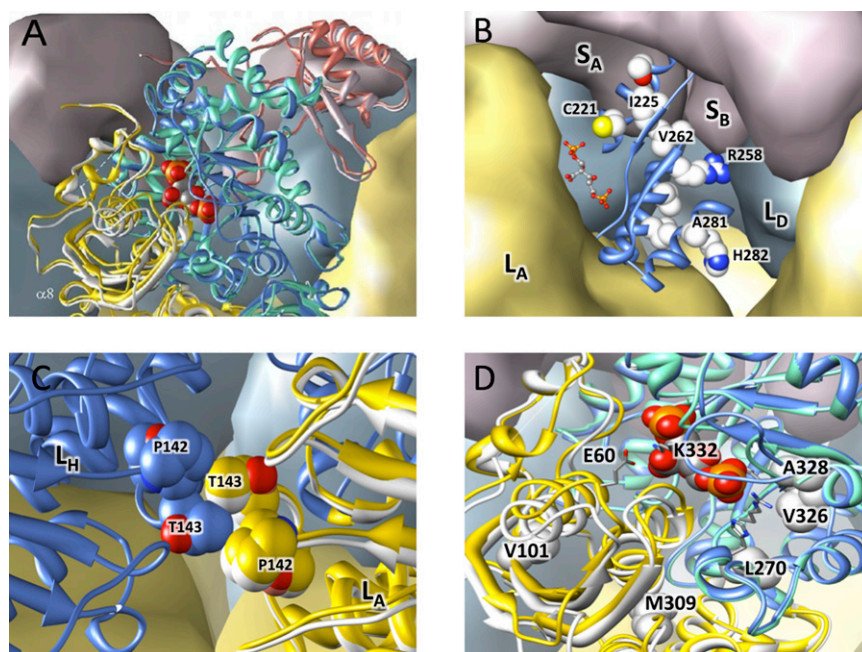


**Fig. S1.** Distribution of stability effects of ancestral and possible mutations. The distribution of stability effects arising from mutations in the evolutionary history of the reconstructed ancestral sequences (black outline) stands in contrast to that of all possible simulated mutations (white outline). These data are the same as in Fig. 3 but here visualized as a histogram.

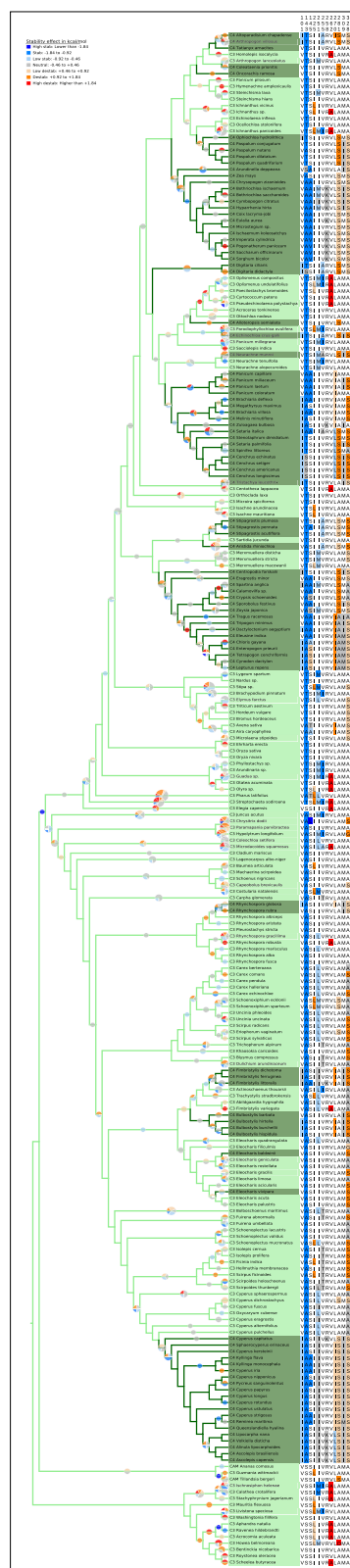




**Fig. S2.** The RubisCO hexadecamer interfaces. Pairs of large subunits form dimers with an extensive interface; four of these dimers form an octomeric ring. The interdimer interfaces are comparatively small, and the overall structure is stabilized by the binding of eight small subunits that bridge dimers. Each large subunit has interfaces with several neighbors. Interfaces are annotated following van Lun et al. (4): LL1 interfaces within the dimers ( $L_A L_B$ ,  $L_C L_D$ ,  $L_E L_F$ , and  $L_G L_H$ ) are shown by dashed gray lines, LL2 ( $L_B L_C$ ,  $L_D L_E$ ,  $L_F L_G$ , and  $L_A L_H$ ) interfaces are in red, and LL3 ( $L_A L_C$ ,  $L_B L_D$ ,  $L_C L_E$ ,  $L_D L_F$ ,  $L_E L_G$ ,  $L_F L_H$ ,  $L_A L_G$ , and  $L_B L_H$ ) interfaces are in blue. LS interfaces of a large subunit with only a small subunit are in green.

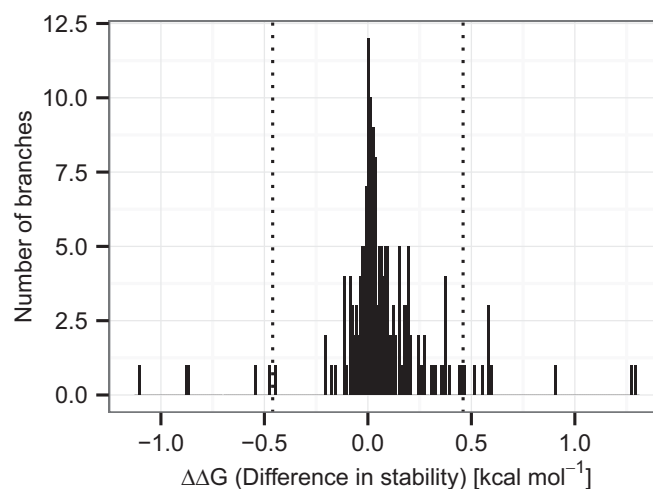


**Fig. S3.** Structural contexts of positively selected sites. On binding RBP and CO<sub>2</sub>, RubisCO undergoes a transition from a relatively disordered open state to an order closed state in which loop 6 (residues 328–337) and the C-terminal tail of the large subunit that are mobile in the open state take up fixed positions that cover the active site. As part of this closure, the N- and C-terminal domains of each subunit and the two subunits of the dimer move relative to each other. (A) Superposition of the C-terminal domain (residues 155–462) of one subunit of rice RubisCO in an open state (3AXM: cyan and white) on the equivalent C-terminal domain in closed form (1WDD: blue and gold) shows that the C-terminal domain itself is altered very little on binding RBP (C $\alpha$  RMSD 0.39 Å) and that the C-terminal domain of the other subunit undergoes a small rotation that results in a more compact structure for the dimer. The rotation is effectively hinged around regions near M390 and G245 and results in reduction of distance between the K450 residues at the top and bottom of the dimer of 1.5 Å. The N-terminal domain of the other subunit of the dimer undergoes greater relative movement (E60 moves 2.8 Å toward the ligand). (B) Within the rigid core of the C-terminal domain of each large subunit (exemplified here by L<sub>B</sub>), a cluster of residues under positive selection are found at the L<sub>B</sub>S<sub>B</sub> interface (and to a lesser extent the L<sub>B</sub>L<sub>D</sub> interface) and at the base of the  $\alpha/\beta$  barrel (in a contiguous stretch of sequence from 221 to 282). (C) In the closed form, residues P142 and T143 on  $\alpha$ -helix 8 of L<sub>A</sub> (gold) intercalate with their counterparts in the neighboring lower left N-terminal domain of L<sub>H</sub> (blue). With L<sub>A</sub> in the open state (white), a 0.5-Å movement of this helix creates a steric clash between T143C $\gamma$  and the proline of the other subunit. (D) Positively selected residues are found near H298 and H327 in the binding site for the phosphate P5 of RBP near loop 6, in contact with the base of  $\alpha$ -helix 8 on the opposite side from E60, which is responsible for positioning the catalytic K332, and at the subunit and domain interface (M309).



**Fig. S4.** Evolutionary tree with stability effect of mutation. Phylogenetic tree of 240 sequences of Rubisco enzymes. Terminal nodes and branches of  $C_3$  and CAM forms are in light green, whereas  $C_4$  forms are in dark green. Mutations are plotted as a pie chart at each node. The size of the chart is proportional to the number of mutations. The slices are colored according to the stability effect: highly stabilizing (lower than  $-1.84$  kcal/mol, in dark blue), stabilizing ( $-1.84$  to  $-0.92$  kcal/mol, in blue), slightly stabilizing ( $-0.92$  to  $-0.46$  kcal/mol, in light blue), no effect ( $-0.46$  to  $+0.46$  kcal/mol, in gray), slightly destabilizing ( $+0.46$  to  $+0.92$  kcal/mol, in light orange), destabilizing ( $+0.92$  to  $+1.84$  kcal/mol, orange), and highly destabilizing (higher than  $+1.84$  kcal/mol, in red). The tree is displayed with EvolView (33). Amino acids observed at sites under strong positive selection (TDG09 FDR  $< 1\%$ ) are plotted after each species name. The colors represent the sum of the stability effect from the root to the terminal nodes.





**Fig. S5.** Analysis of the additivity of in-branch mutations. Some branches of the phylogenetic tree of monocot species contain multiple mutations, and hypothetically, their order of mutation may affect their contribution to stability. It is only practical to treat such mutations as independent and assume that their individual effects are additive. To test the validity of this assumption, we compute the differences of the sums of the calculated  $\Delta\Delta G_{\text{fold}}$  of individual mutations and the  $\Delta\Delta G_{\text{fold}}$  of all mutations made simultaneously for those branches with multiple mutations. The histogram of these differences shows that the stability effects of in-branch mutations are approximately additive, and only in rare cases does the difference exceed the accuracy of FoldX results (dashed vertical lines at  $\pm 0.46$  kcal/mol).

## Other Supporting Information Files

[Table S1 \(DOC\)](#)

[Table S2 \(DOC\)](#)

[Table S3 \(DOC\)](#)

[Table S4 \(DOC\)](#)